Electron. Commun. Probab. **31** (2026), article no. 55, 1–15.
https://doi.org/10.1214/26-ECP804
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

Limit theorems for stochastic gradient descent in high-dimensional single-layer networks*

Parsa Rangriz[†‡]

Abstract

This paper studies the high-dimensional scaling limits of online stochastic gradient descent (SGD). Building on the work of Ben Arous, Gheissari, and Jagannath on the effective dynamics of SGD, we study the critical scaling regime of the step size for single-layer networks. Below this regime the effective dynamics are governed by deterministic (ballistic) limits, whereas at the critical scale a correction term emerges that changes the phase diagram. Near the fixed points of these dynamics, we show that the diffusive (SDE) limit of the rescaled correlation is an Ornstein–Uhlenbeck process. More precisely, it is mean-reverting whenever the information exponent is at least three. At information exponent two the drift has no universal sign, and the fixed point may become repelling; we show this explicitly for phase retrieval, where the sign is determined by the step size and the noise level. These results illustrate the limitations of deterministic scaling limits in capturing stochastic fluctuations in high-dimensional learning dynamics.

Keywords: stochastic gradient descent; functional central limit theorem; single-index models.
MSC2020 subject classifications: 60F17.
Submitted to ECP on November 6, 2025, final version accepted on August 16, 2026.
Supersedes arXiv:2511.02258.

1 Introduction

Stochastic gradient descent (SGD) is one of the most widely used algorithms in machine learning, optimization, and data science. Since its introduction by Robbins and Monro in the 1950s [24], a main challenge in the theory of machine learning has been to understand how SGD navigates the non-convex loss landscape to train neural networks. While early works focused on fixed-dimensional settings, e.g., [21, 5, 6, 12, 15, 19], recent work has shifted toward high-dimensional regimes, where both the sample size and parameter dimension grow, e.g., [23, 16, 20, 14, 28].

In fixed dimensions, the behavior of SGD can be studied using classical asymptotic methods and stochastic approximation theory, introduced by McLeish (1976) [21], including pathwise limit theorems such as the functional central limit theorem (FCLT) and

*This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Cette recherche a été enterprise grâce, en partie, au soutien financier du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [RGPIN-2020-04597].

[†]Department of Mathematics, University of California San Diego, USA.
E-mail: prangriz@ucsd.edu

[‡]This work was done while the author was affiliated with the Department of Statistics and Actuarial Science, University of Waterloo, Canada.

large deviation principles. In the regime of sufficiently small step size (learning rate) of the algorithm, with a fixed loss function, the scaling limit of the SGD trajectory has been shown to converge to the solution of a gradient flow problem, e.g., [2, 10, 22, 25, 27]. There has been also growing interest in higher-order works via diffusion approximations, including asymptotic expansions of the SGD trajectory in terms of the step size, e.g., [1, 16, 17, 18].

By contrast, in high-dimensional settings, tracking the full SGD trajectory is often infeasible. A common approach is to analyze scaling limits of lower-dimensional summary statistics under regularity and simplifying assumptions. A major development in this direction came in the late 1990s, when Saad and Solla [26, 8] introduced order parameters, such as the overlap matrix in the single-index model, drawing inspiration from the statistical physics of spin glasses. This perspective, known as dynamical mean-field theory (DMFT), has since provided a powerful framework for analyzing the high-dimensional dynamics of SGD, e.g., [7, 13, 9, 29, 11].

Building on this breakthrough, subsequent research has focused on characterizing the classes of functions that SGD can efficiently learn, in terms of both time and sample complexity. The dynamics are typically studied in the ballistic phase, where summary statistics evolve on a macroscopic scale and are well-approximated by an ordinary differential equation (ODE). In this regime, the macroscopic behavior follows a deterministic scaling limit akin to the population gradient flow, e.g., [13, 29, 11].

In single-index models, for example, the time required for learning scales with the dimension, which depends sensitively on the geometry of the loss landscape. To analyze the dynamics near fixed points of these ODEs, Ben Arous, Gheissari, and Jagannath [3] introduced the concept of the information exponent, a geometric quantity that captures how SGD explores the loss landscape. They showed that there are three distinct behaviors in which the time to weakly recover is linear, quasi-linear, or polynomial, depending on whether the information exponent is less than two, equal to two, or greater than two, respectively.

In this paper, we establish a FCLT for the rescaled dynamics of SGD in single-index models. Specifically, we analyze the diffusive phase, where the summary statistics fluctuate microscopically around fixed points, and the ballistic approximation breaks down. In a critical scaling regime for the step size, an additional correction term arises in the dynamics, leading to significant deviations from the population gradient flow. In microscopic neighborhoods of a fixed point, the effective dynamics become stochastic and are governed by stochastic differential equations (SDEs), which may display a wide range of behaviors, including degenerate cases.

Our main focus is on learning single-index models whose activation has information exponent at least two. We show that, under random initialization, the high-dimensional trajectory of SGD with stochastic corrections deviates from the deterministic limit with no population corrector predicted by DMFT. At the critical step size the correlation stays at zero along the ballistic limit, so the recovery behavior is visible only in its rescaled fluctuations. The dynamics converge to an Ornstein–Uhlenbeck (OU) process, and the information exponent decides the sign of its drift: mean-reverting when the exponent is at least three, of either sign when it equals two. We illustrate this for phase retrieval.

2 Setting and assumptions

Suppose that we are given a parametric family of distributions, $(\mathbb{P}_x)_{x\in\mathbb{R}^N}$. According to the teacher-student scenario, the teacher begins by generating a hidden vector $x^* \in \mathbb{R}^N$ from a known prior distribution. Based on a statistical model, the teacher then produces a sequence of i.i.d. observations $(Y_k)_k$, each generated conditionally on x^*

and parameterized by elements in $\mathcal{Y}_N \subseteq \mathbb{R}^N$, from $P_N = \mathbb{P}_{x^*}$, which we call the *data distribution*. The number of observations is indexed by $k \in \{1, \dots, N\}$. The teacher then provides the dataset $(Y_k)_k$, along with partial information about the generative model, to the student. The student's objective is to infer the hidden variables x^* using only the observed data and the provided model information.

Suppose a sequence of parameter iterates $(X_k)_k$ lies in a high-dimensional space $\mathcal{X}_N \subseteq \mathbb{R}^N$, and the training data $(Y_k)_k$ takes values in $\mathcal{Y}_N \subseteq \mathbb{R}^N$. The learning proceeds via online SGD with respect to a loss function $L_N : \mathcal{X}_N \times \mathcal{Y}_N \to \mathbb{R}$, and a constant step-size $\delta_N = c_\delta / N$ (for a fixed constant $c_\delta > 0$ independent of N), according to the update rule

$$X_{k+1} = X_k - \delta_N \nabla L_N(X_k; Y_{k+1}),$$

initialized with a random vector $X_0 \sim \mu_N \in \mathcal{M}_1(\mathbb{R}^N)$, where $\mathcal{M}_1(\mathbb{R})$ denotes the space of probability measures on $\mathbb{R}$. Our goal is to understand the evolution of the sequence (X_k) in the high-dimensional limit as $N \to \infty$. To this end, suppose that we are given a sequence of functions $\mathbf{u}_N \in C^1(\mathbb{R}^N; \mathbb{R}^l)$ for some fixed l, where $\mathbf{u}_N(x) = (u_1^N(x), \dots, u_l^N(x))$, and more precisely, our goal is to understand the evolution of $\mathbf{u}_N(X)$.

To develop a scaling limit, we need some regularity assumptions on the relationship between how the step-size δ_N scales in relation to the loss L_N, its gradients, and the data distribution P_N. Therefore, we define the *sample-wise error* as follows

$$H(x, Y) := L_N(x, Y) - \Phi(x) \quad \text{where} \quad \Phi(x) := \mathbb{E}[L_N(x, Y)].$$

In the following, we suppress the dependence of H on Y and instead view H as a random function of x, which we denote as $H(x)$. We let $V(x) = \mathbb{E}[\nabla H(x) \otimes \nabla H(x)]$ be the covariance matrix for ∇H at x.

Consider the following model of supervised learning with a *single-layer network*[1]: Suppose we are given a (possibly) non-linear activation function $f : \mathbb{R} \to \mathbb{R}$, a set of feature vectors $(a_k)_k$, and additive noisy responses $(\epsilon_k)_k$ of the form

$$y_k = f(\langle a_k, x^* \rangle) + \epsilon_k,$$

and for the sake of simplicity, we consider quadratic loss functions

$$L_N(x, Y) = L_N(x, (y, a)) = (y - f(\langle a, x \rangle))^2.$$

Let us focus on the most studied regime, namely where $(a_k)_k$ are i.i.d. standard Gaussian vectors in $\mathbb{R}^N$; for the $(\epsilon_k)_k$ we only assume they are i.i.d. mean zero with variance C_ϵ and finite $4 + \delta$-th moment for some $\delta > 0$.

Note that we may write the population loss as

$$\Phi(x) = \mathbb{E}\left[(f(\langle a, x \rangle) - f(\langle a, x^* \rangle))^2 \right] + C_\epsilon.$$

Also, in our case, the sufficient number of summary statistics we need for the single-index model is two, and let $x^* \in \mathbb{S}^{N-1}$ be a fixed unit vector. Now, we define the following summary statistics $\mathbf{u}_N(x) = (u_1^N(x), u_2^N(x))$

$$u_1^N(x) := m(x) = \langle x, x^* \rangle, \qquad u_2^N(x) := r_\perp^2(x) = \|x\|^2 - m^2(x),$$

where the loss distribution only depends on $(m, r_\perp^2)$ and the population loss is of the form $\Phi(x) = \phi(m, r_\perp^2)$. We call m the *correlation* of x with x^*. We also refer to $r_\perp > 0$ as the *radius*.

[1] This model and special cases thereof have been studied under many different names by a broad range of communities: single-layer neural networks, teacher-student networks, and single-index models.

Remark 2.1. Note that the model is unchanged by the substitution $f \mapsto f + c$ for any constant $c \in \mathbb{R}$. If the teacher's activation is shifted, the observed responses become $y_k = (f + c)(\langle a_k, x^* \rangle) + \epsilon_k$, and the student's loss $L_N(x, Y)$ under the shift is invariant. Consequently, the loss L_N, the population loss Φ, the sample-wise error H, and the covariance V are invariant under the shift.

To ensure the tightness of the trajectories of the summary statistics, [4] impose two key assumptions, namely *localizablity* and *asymptotic closability* on the triplet $(\mathbf{u}_N, L_N, P_N)$ and the learning rate δ_N.

Definition 2.2. *A triple $(\mathbf{u}_N, L_N, P_N)$ is δ_N-localizable with localizing sequence $(E_K)_K$ if there is an exhaustion by compacts $(E_K)_K$ and constant C_K (independent of N) such that*

1. $\max_i \sup_{x \in \mathbf{u}_N^{-1}(E_K)} \|\nabla^2 u_i^N\|_{op} \le C_K \delta_N^{-1/2}$, *and* $\max_i \sup_{x \in \mathbf{u}_N^{-1}(E_K)} \|\nabla^3 u_i^N\|_{op} \le C_K$.
2. $\sup_{x \in \mathbf{u}_N^{-1}(E_K)} \|\nabla \Phi\| \le C_K$ *and* $\sup_{x \in \mathbf{u}_N^{-1}(E_K)} \mathbb{E}[\|\nabla H\|^8] \le C_K \delta_N^{-4}$.
3. $\max_i \sup_{x \in \mathbf{u}_N^{-1}(E_K)} \mathbb{E}[\langle \nabla H, \nabla u_i^N \rangle^4] \le C_K \delta_N^{-2}$, *and*
   $\max_i \sup_{x \in \mathbf{u}_N^{-1}(E_K)} \mathbb{E}[\langle \nabla^2 u_i^N, \nabla H \otimes \nabla H - V \rangle^2] = o(\delta_N^{-3})$

Definition 2.3. *A family of summary statistics $(\mathbf{u}_N)$ are asymptotically closable for learning rate δ_N if $(\mathbf{u}_N, L_N, H)$ are δ_N-localizable with localizing sequence $(E_K)_K$, and furthermore there exist locally Lipschitz functions $\mathcal{H} : \mathbb{R}^l \to \mathbb{R}^l$ and $\Sigma : \mathbb{R}^l \to \mathbb{R}^{l \times l}$, such that*

$$\sup_{x \in \mathbf{u}_N^{-1}(E_k)} \|(-\mathcal{A}_N + \delta_N \mathcal{L}_N)\mathbf{u}_N(x) - \mathcal{H}(\mathbf{u}_N(x))\| \to 0,$$

and

$$\sup_{x \in \mathbf{u}_N^{-1}(E_k)} \|\delta_N J_N V J_N^T - \Sigma(\mathbf{u}_N(x))\| \to 0$$

where $\mathcal{A}_N = \langle \nabla \Phi, \nabla \rangle$ and $\mathcal{L}_N = \frac{1}{2}\langle V, \nabla^2 \rangle$. We call $\mathcal{H}$ the effective drift, and Σ the effective volatility.

For the sake of simplicity, suppose that not only the asymptotic closability is satisfied, but each of the two terms $\mathcal{A}_N \mathbf{u}_N$ and $\delta_N \mathcal{L}_N \mathbf{u}_N$ in Definition 2.3 individually admit $N \to \infty$ limits: namely there exist $\mathcal{F}, \mathcal{G} : \mathbb{R}^l \to \mathbb{R}^l$ such that

$$\sup_{x \in \mathbf{u}_N^{-1}(E_K)} \|\mathcal{A}_N \mathbf{u}_N(x) - \mathcal{F}(\mathbf{u}_N(x))\| \to 0,$$

and

$$\sup_{x \in \mathbf{u}_N^{-1}(E_K)} \|\delta_N \mathcal{L}_N \mathbf{u}_N(x) - \mathcal{G}(\mathbf{u}_N(x))\| \to 0,$$

evidently $\mathcal{H} = -\mathcal{F} + \mathcal{G}$, and we call $\mathcal{F}$ and $\mathcal{G}$ the *population drift* and the *population corrector*, respectively.

Then the corresponding (possibly stochastic) differential equation of online SGD [4] is given by

$$\mathrm{d}\mathbf{u}_t = (-\mathcal{F}(\mathbf{u}_t) + \mathcal{G}(\mathbf{u}_t))\mathrm{d}t + \sqrt{\Sigma(\mathbf{u}_t)}\mathrm{d}\mathbf{B}_t, \tag{2.1}$$

where $\mathbf{B}_t$ is a standard Brownian motion in $\mathbb{R}^l$.

Recall that the Hermite polynomials, which we denote by $(h_k(x))_{k=0}^{\infty}$, are the normalized orthogonal polynomials of the Gaussian distribution $\gamma(x) \propto \exp(-x^2/2)dx$. Define the k-th Hermite coefficient for an activation function $f \in L^2(\gamma)$ by

$$\alpha_k := \langle f, h_k \rangle_{L^2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(z) h_k(z) e^{-z^2/2} dz.$$

Also, we define the norm $\|f\|_{L^2}^2 := \langle f, f\rangle_{L^2}$. As long as f' has at most polynomial growth, the population loss is differentiable, and 2.1 exists.

According to [3], under the regularity conditions previously discussed, in the following we define a key quantity governing the performance of online SGD.

Definition 2.4. *We say that an activation function $f : \mathbb{R} \to \mathbb{R}$ has information exponent k if the first non-zero coefficient in its Hermite expansion is the kth coefficient, i.e., $\alpha_i = 0$ for all $1 \le i < k$ and $\alpha_k \neq 0$.*

3 Main results

In this work, we focus exclusively on activation functions with an information exponent of at least two. This choice implies that the corresponding sample complexity of online SGD grows at least quasi-linearly—or potentially polynomially—with the dimension, placing our analysis squarely in the more challenging high-dimensional learning regime.

Even with this relative simplicity, we encounter various ODE and SDE limits following the general form of (2.1). Indeed, we find dynamical phase transitions corresponding to the aforementioned threshold in our model. Our analysis focuses exclusively on the most interesting, critical step-size scaling $\delta_N = c_\delta/N$ corresponding to the proportional asymptotics regime from the random matrix theory literature.

Before stating our main results, we introduce a notation for the Gaussian functionals of f that recur throughout, and the mentioned regularity assumptions on f. For $r_\perp > 0$ and $a \sim \mathcal{N}(0,1)$, write

$$\begin{aligned} \mathsf{a}(r_\perp) &:= \mathbb{E}[f'^2(ar_\perp)], \qquad \mathsf{b}(r_\perp) := \mathbb{E}[f(ar_\perp)f''(ar_\perp)], \qquad \mathsf{s}(r_\perp) := \mathbb{E}[f''(ar_\perp)], \\ \mathsf{t}(r_\perp) &:= \mathbb{E}[f'^2(ar_\perp)f(ar_\perp)], \qquad \mathsf{q}(r_\perp) := \mathbb{E}[f'^2(ar_\perp)f^2(ar_\perp)]. \end{aligned} \tag{3.1}$$

Assumption 3.1. *Assume that the activation function f holds the following conditions:*

(A1) $f \in L^2(\gamma)$ is twice-differentiable a.e. and f', f'' have at most polynomial growth;
(A2) the information exponent of f is at least two.

3.1 Sub-critical scaling regime

We are now ready to present our main results.

Theorem 3.2. *Let $(X_k^{\delta_N})_k$ be SGD initialized from $X_0 \sim \mu_N$ for $\mu_N \in \mathcal{M}_1(\mathbb{R}^N)$ with the learning rate δ_N for the quadratic loss L_N of a single-index model. Suppose that the activation function f satisfies Assumption 3.1. For the corresponding summary statistics (correlation and radius) $\mathbf{u}_N = (u_i^N)_{i=1}^2 = (m, r_\perp^2)$, let $(\mathbf{u}_N(t))_t$ be the linear interpolation of $(\mathbf{u}_N(X^{\delta_N}_{\lfloor t\delta_N^{-1}\rfloor}))_t$. Then $\mathbf{u}_N$ are asymptotically closable with learning rate $\delta_N = c_\delta/N$. Moreover, $\mathbf{u}_N = (m, r_\perp^2)$ converges as $N \to \infty$ to the solution of the following ODE initialized from the pushforward of the initial data $\lim_{N\to\infty}(\mathbf{u}_N)_*\mu_N$,*

$$\begin{cases} \dfrac{\mathrm{d}m}{\mathrm{d}t} = -2\mathbb{E}_{a_1,a_2}[a_1 f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))], \\ \dfrac{\mathrm{d}r_\perp^2}{\mathrm{d}t} = -4\mathbb{E}_{a_1,a_2}[a_2 r_\perp f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))] \\ \qquad\qquad +4c_\delta \mathbb{E}_{a_1,a_2}[f'^2(a_1 m + a_2 r_\perp)((f(a_1 m + a_2 r_\perp) - f(a_1))^2 + C_\epsilon)], \end{cases} \tag{3.2}$$

where a_1, a_2 are i.i.d. standard Gaussian variables.

We can obtain the following result when we restrict the initialization of the algorithm to be chosen randomly from a gaussian distribution, i.e., $X_0 \sim \mathcal{N}(0, \frac{\sigma^2}{N} I_N)$, then $(\mathbf{u}_N)_*\mu_N \to \delta_{(0,\sigma^2)}$ weakly for some fixed σ^2.

Corollary 3.3. *Under the conditions of Theorem 3.2, $r_\perp^2$ converges as $N \to \infty$ to the solution of the following ODE initialized from the pushforward of the initial data $\lim_{N\to\infty}(\mathbf{u}_N)_*\mu_N = \delta_{(0,\sigma^2)}$,*

$$\frac{\mathrm{d}r_\perp^2}{\mathrm{d}t} = -4r_\perp^2\,\gamma(r_\perp) + 4c_\delta\,\mathsf{n}(r_\perp), \tag{3.3}$$

where

$$\begin{cases} \gamma(r_\perp) = \mathsf{a}(r_\perp) + \mathsf{b}(r_\perp) - \alpha_0\mathsf{s}(r_\perp) = \mathbb{E}_{a_2}[f'^2(a_2r_\perp) + (f(a_2r_\perp) - \alpha_0)f''(a_2r_\perp)] \\ \mathsf{n}(r_\perp) := \mathbb{E}_{a_2}\big[f'^2(a_2r_\perp)\big(f(a_2r_\perp) - \alpha_0\big)^2\big] + \mathsf{a}(r_\perp)\big(C_\epsilon + \|f - \alpha_0\|_{L^2}^2\big) \;>\; 0. \end{cases}$$

Moreover, $m \equiv 0$ at all time, where the functionals $\mathsf{a}, \mathsf{b}, \mathsf{s}$ are as in (3.1).

Remark 3.4. Whether the ODE (3.3) admits a fixed point depends jointly on the activation f and on the step-size constant c_δ. The first term of (3.3) is the confinement produced by the population drift $\mathcal{F}_{r_\perp^2}$, while the second is the heating produced by the corrector $\mathcal{G}_{r_\perp^2}$; only the latter carries c_δ. Since $\mathsf{n}(r_\perp) > 0$, a radius $r_\perp^* > 0$ is a fixed point if and only if $c_\delta = \Psi(r_\perp^*)$, where

$$\Psi(r_\perp) := \frac{r_\perp^2\,\gamma(r_\perp)}{\mathsf{n}(r_\perp)}.$$

Hence a fixed point exists precisely when c_δ lies in the range of Ψ, and in particular only when $c_\delta \le c_\delta^{\mathrm{crit}} := \sup_{r_\perp>0}\Psi(r_\perp)$; above this threshold the heating dominates at every radius and $r_\perp^2$ diverges.

3.2 Critical scaling regime

Let us consider a rescaling regime of $\mathbf{u}_N$ in a microscopic neighborhood of the fixed point $m = 0$. This captures the initial phase from a random start if $\mu_N \sim \mathcal{N}(0, \frac{\sigma^2}{N} I_N)$ for some fixed $\sigma^2 > 0$. Then the pushforward satisfies then $(\mathbf{u}_N)_*\mu_N \to \delta_{(0,\sigma^2)}$ weakly. Now rescale and let $\tilde{\mathbf{u}}_N = (\tilde{m}, r_\perp^2)$ where $\tilde{m} = \sqrt{N}m$. Evidently, $\bar{\nu} = \lim_{N\to\infty}(\tilde{\mathbf{u}}_N)_*\mu_N = \mathcal{N}(0,\sigma^2)\otimes\delta_{\sigma^2}$.

Theorem 3.5. *Let $(X_k^{\delta_N})_k$ be SGD initialized from $X_0 \sim \mathcal{N}(0, \frac{\sigma^2}{N} I_N)$ for some fixed σ^2 with learning rate δ_N for the quadratic loss L_N of a single-index model. Suppose that the activation function f satisfies Assumption 3.1. For the corresponding summary statistics (rescaled correlation and radius) $\tilde{\mathbf{u}}_N = (\tilde{u}_i^N)_{i=1}^2 = (\tilde{m}, r_\perp^2)$, let $(\tilde{\mathbf{u}}_N(t))_t$ be the linear interpolation of $(\tilde{\mathbf{u}}_N(X^{\delta_N}_{\lfloor t\delta_N^{-1}\rfloor}))_t$. Then $\tilde{\mathbf{u}}_N$ are asymptotically closable with learning rate $\delta_N = c_\delta/N$. Moreover, $\tilde{\mathbf{u}}_N = (\tilde{m}, r_\perp^2)$ converges as $N \to \infty$ to the solution of the following SDE initialized from $\mathcal{N}(0,\sigma^2)\otimes\delta_{\sigma^2}$,*

$$\begin{cases} \mathrm{d}\tilde{m} = -2\tilde{m}\,\big(\gamma(r_\perp) - \alpha_2\mathsf{s}(r_\perp)\big)\mathrm{d}t + 2\sqrt{c_\delta\big(\mathsf{q}(r_\perp) - 2(\alpha_0+\alpha_2)\mathsf{t}(r_\perp) + \mathsf{a}(r_\perp)\eta\big)}\,\mathrm{d}B_t, \\ \dfrac{\mathrm{d}r_\perp^2}{\mathrm{d}t} = -4r_\perp^2\,\gamma(r_\perp) + 4c_\delta\,\mathsf{n}(r_\perp), \end{cases} \tag{3.4}$$

where

$$\begin{cases} \gamma(r_\perp) = \mathsf{a}(r_\perp) + \mathsf{b}(r_\perp) - \alpha_0\mathsf{s}(r_\perp) = \mathbb{E}_{a_2}[f'^2(a_2r_\perp) + (f(a_2r_\perp) - \alpha_0)f''(a_2r_\perp)] \\ \eta = \|f\|_{L^2}^2 + 2\|f'\|_{L^2}^2 + 2\langle f, f''\rangle_{L^2} + C_\epsilon, \\ \mathsf{n}(r_\perp) := \mathbb{E}_{a_2}\big[f'^2(a_2r_\perp)\big(f(a_2r_\perp) - \alpha_0\big)^2\big] + \mathsf{a}(r_\perp)\big(C_\epsilon + \|f - \alpha_0\|_{L^2}^2\big) \end{cases}$$

with a_1, a_2 i.i.d. standard Gaussian, $\alpha_0 = \mathbb{E}_{a_1}[f(a_1)]$, and $\alpha_2 = \mathbb{E}_{a_1}[f''(a_1)]$ and the functionals $\mathsf{a}, \mathsf{b}, \mathsf{s}, \mathsf{t}, \mathsf{q}$ are as in (3.1).

By Remark 2.1, since the SGD trajectory $(X_k)_k$ is driven through ∇L_N, every limiting ODE and SDE derived above is therefore obligated to be invariant as well.

Now, if the information exponent is at least three and the radius term of the rescaled summary statistics starts from the initial point $r_\perp^*$, then a *mean-reverting* OU process appears.

Corollary 3.6. *Assume the conditions of Theorem 3.5 hold, except that the information exponent of f is now at least three, i.e., $\alpha_2 = 0$. If $\mu_N \sim \mathcal{N}(0, \frac{{r_\perp^*}^2}{N} I_N)$ where ${r_\perp^*}^2$ is the fixed-point of the ODE for $r_\perp^2$ of (3.4), then $\tilde{m}$ converges as $N \to \infty$ to the solution of the following mean-reverting OU process, i.e., $\gamma(r_\perp^*) > 0$,*

$$\mathrm{d}\tilde{m} = -2\tilde{m}\,\gamma(r_\perp^*)\mathrm{d}t + 2\sqrt{c_\delta\big(\mathsf{q}(r_\perp^*) - 2\alpha_0\mathsf{t}(r_\perp^*) + \mathsf{a}(r_\perp^*)\eta\big)}\,\mathrm{d}B_t. \tag{3.5}$$

Remark 3.7. The mean-reverting OU process in Corollary 3.6 is special to information exponent at least three, and does not necessarily hold at information exponent exactly two. If $\alpha_2 \neq 0$, the drift coefficient becomes $\theta(r_\perp^*) := \gamma(r_\perp^*) - \alpha_2\mathsf{s}(r_\perp^*)$. Then, there is a competition between $\gamma(r_\perp^*)$ and $\alpha_2\mathsf{s}(r_\perp^*)$, i.e., the fixed point $m = 0$ can be mean-reverting or mean-repelling depending on f, with no universal sign.

3.3 Example: Phase retrieval

Phase retrieval, $f(x) = x^2$, makes the dichotomy of Remark 3.7 completely explicit, exhibiting both signs of the drift within a single model of information exponent two. By defining $u = r_\perp^2$, one gets $\Psi(u) = (3u-1)/\big(2(15u^2 - 6u + 3 + C_\epsilon)\big)$, whence

$$c_\delta^{\mathrm{crit}} = \frac{\sqrt{5(8+3C_\epsilon)} - 2}{4(12+5C_\epsilon)},$$

equal to $(\sqrt{10}-1)/24 \approx 0.0901$ when $C_\epsilon = 0$. Below the threshold the two roots of $\Psi(u) = c_\delta$ are a stable fixed point and an unstable barrier; as $c_\delta \downarrow 0$ the former decreases to the population value $u = 1/3$ and the latter escapes to infinity, while at $c_\delta = c_\delta^{\mathrm{crit}}$ they merge in a saddle-node bifurcation.

Here $\alpha_0 = 1$, $\alpha_2 = 2$. Also, $\mathsf{a}(r_\perp) = 4r_\perp^2$, $\mathsf{b}(r_\perp) = 2r_\perp^2$, $\mathsf{s}(r_\perp) = 2$, $\mathsf{t}(r_\perp) = 12r_\perp^4$, $\mathsf{q}(r_\perp) = 60r_\perp^6$, so that $\gamma(r_\perp) = 2(3r_\perp^2 - 1)$ while the drift of the rescaled correlation is governed by

$$\theta(r_\perp) = \gamma(r_\perp) - \alpha_2\mathsf{s}(r_\perp) = 6\big(r_\perp^2 - 1\big). \tag{3.6}$$

Writing $u = r_\perp^2$, the RHS of (3.3) becomes $4u\big(2 - 6u + 4c_\delta(15u^2 - 6u + 3 + C_\epsilon)\big)$, whose positive zeros are

$$u_\pm = \frac{(24c_\delta + 6) \pm \sqrt{36 - 192c_\delta - (2304 + 960C_\epsilon)c_\delta^2}}{120\,c_\delta},$$

with u_- stable and u_+ an unstable barrier. Initializing at $r_\perp^* = \sqrt{u_-}$, the rescaled correlation is then an OU process,

$$\mathrm{d}\tilde{m} = -12\big({r_\perp^*}^2 - 1\big)\,\tilde{m}\,\mathrm{d}t + 4r_\perp^*\sqrt{c_\delta\big(15{r_\perp^*}^4 - 18{r_\perp^*}^2 + 15 + C_\epsilon\big)}\,\mathrm{d}B_t, \tag{3.7}$$

whose volatility never degenerates, as $15v^2 - 18v + 15 + C_\epsilon > 0$ for all $v \geq 0$. By (3.6) the process is mean-reverting when $r_\perp^* > 1$, and evaluating the quadratic above at $u = 1$ shows that this happens if and only if

$$C_\epsilon > 4 \qquad \text{and} \qquad \frac{1}{12 + C_\epsilon} < c_\delta \leq c_\delta^{\mathrm{crit}} = \frac{\sqrt{5(8+3C_\epsilon)} - 2}{4(12+5C_\epsilon)}.$$

The two thresholds coincide at $C_\epsilon = 4$, where $c_\delta^{\mathrm{crit}} = 1/16$. Hence for $C_\epsilon \le 4$ the correlation is mean-*repelling* at the stable radius for every admissible step size, and in particular $\theta(r_\perp^*) \to -4$ as $c_\delta \downarrow 0$, since $u_- \downarrow 1/3$; only in a narrow window of large step sizes and large noise does the fixed point become mean-reverting. This is precisely the behaviour that Corollary 3.6 rules out at information exponent at least three, and it is the single-index analogue of the signal-to-noise dependent drift found for matrix PCA in [4].

4 Discussion

We showed that, with random initialization, the effective dynamics of SGD deviate from the trajectories predicted by deterministic limits in high-dimensional settings. As discussed earlier, DMFT describes a deterministic scaling limit via an ODE approximation akin to the population gradient flow. However, in the critical step-size regime, diffusive effects emerge, and the effective dynamics depart from this deterministic description due to the presence of a stochastic correction, which we refer to as the population corrector. At the critical step-size $\delta_N = c_\delta/N$, and when the information exponent is at least two, the correlation $m = \langle x, x^* \rangle$ stays at zero, which is a fixed point of the corresponding ballistic limit. This indicates that deterministic limits fail to fully capture the recovery behavior in this regime.

Near a fixed point of the radius equation, we showed that the rescaled correlation $\tilde{m}$ follows an OU process, and the information exponent decides whether this process pulls $\tilde{m}$ back to zero or pushes it away. If the information exponent is at least three, the drift always points back to zero, so $m = 0$ stays stable even once the stochastic correction is taken into account. If it is exactly two, the drift may point either way, depending on the activation function, and the fixed point can then be mean-repelling rather than mean-reverting.

As an example, we investigated phase retrieval, which shows how much the corrector can change the dynamics. Its population drift is by itself confining: this drift vanishes at $r_\perp^2 = 1/3$, so the gradient flow of Φ alone would leave the radius bounded. The heating comes entirely from the corrector, and once c_δ passes c_δ^{crit} this control is lost and the radius grows without bound. The two signs of the drift also match what is known about sample complexity. At information exponent two, a drift that pushes $\tilde{m}$ away from zero lets the small initial correlation, of size $O(N^{-1/2})$, grow. This is why weak recovery needs only quasi-linearly many samples [3]. At information exponent at least three, a drift that pulls $\tilde{m}$ back keeps the correlation trapped near the origin, so polynomially many samples are needed instead.

5 Proofs

5.1 Proofs of Theorem 3.2 and Corollary 3.3

Lemma 5.1. *In a single-index model, the distribution of the loss $L_N(x,(a,y))$ depends only on $\mathbf{u}_N = (m, r_\perp^2)$. Also, $\mathbf{u}_N$ is δ_N-localizable for E_K being the centered balls of radius K in $\mathbb{R}^2$.*

Proof. We check the items in Definition 2.2. By rotational invariance of the Gaussian ensemble, we may take $x^* = v$ where v is the first basis vector of $\mathbb{R}^N$. First note that for every x, since f is differentiable and f' is of at most polynomial growth,

$$\nabla \Phi = \partial_m \phi \nabla m + \partial_{r_\perp^2} \phi \nabla r_\perp^2,$$

where

$$\begin{cases}\partial_m \phi = 2\mathbb{E}_{a_1,a_2}[a_1 f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))], \\ \partial_{r_\perp^2}\phi = \frac{1}{r_\perp}\mathbb{E}_{a_1,a_2}[a_2 f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))].\end{cases}$$

Note that, $(a_k)_{k=1}^N$ are i.i.d Gaussian variables as stated in the Introduction, but here by rotational invariance, we rename a_2 such that $\mathbb{E}[f(\langle a, x\rangle)] = \mathbb{E}[f(a_1 m + a_2 r_\perp)]$. In particular, a_2 here does not correspond to the original second coordinate, but to the component orthogonal to v.

One may express the derivatives for u_N as

$$\nabla m = v, \qquad \nabla r_\perp^2 = 2(x - mv).$$

Notice that $\nabla^2 m = 0$, while $\nabla^2 r_\perp^2 = 2(I - vv^T)$, and $\nabla^l m = \nabla^l r_\perp^2 = 0$ for all $l \geq 3$. It yields that

$$\langle \nabla m, \nabla m\rangle = 1, \qquad \langle \nabla m, \nabla r_\perp^2\rangle = 0, \qquad \langle \nabla r_\perp^2, \nabla r_\perp^2\rangle = 4r_\perp^2.$$

For part (2), one may write

$$\|\nabla \Phi\| \leq |\partial_m \phi| \|\nabla m\| + |\partial_{r_\perp^2}\phi| \|\nabla r_\perp^2\|,$$

the bounding quantity is evidently a continuous function of $m, r_\perp^2$ and therefore as long as x is such that $(m, r_\perp^2) \in E_K$, it is bounded by some constant C_K.

Recall,

$$H(x, a) = (f(\langle a, x\rangle) - f(a_1))^2 - \Phi(x).$$

Then the derivatives of H are given by

$$\nabla H(x, a) = 2a f'(\langle a, x\rangle)(f(\langle a, x\rangle - f(a_1)) - \nabla \Phi(x).$$

Since f' has at most polynomial growth, $\|a\| = O_p(\sqrt{N})$, and $\|\nabla \Phi\| = O_p(1)$, where O_p is stochastic boundedness, we get

$$\|\nabla H\| \leq 2\|a\| \, |f'(\langle a, x\rangle)(f(\langle a, x\rangle) - f(a_1))| + \|\nabla \Phi\| = O_p(\sqrt{N}).$$

Therefore, there exists $C_K(f) > 0$ independent of N such that

$$\mathbb{E}[\|\nabla H\|^8] \leq C_K(f) N^4.$$

Moving on item (3), for every w,

$$\mathbb{E}[\langle \nabla H, w\rangle^4] \leq \mathbb{E}[\|\nabla H\|^4]\|w\|^4 \leq C_K(f) N^2.$$

If $w = \nabla m = v$, then $\|w\| = 1$ and if $w = \nabla r_\perp^2 = 2(x - mv)$ then $\|w\| = 4r_\perp^2 \leq c_K$, for some constant c_k, so in both cases the upper bound is at most $C_K(f)N^2$. Furthermore,

$$\mathbb{E}[\langle \nabla^2 r_\perp^2, \nabla H \otimes \nabla H - V\rangle^2] \leq 4\mathbb{E}[\langle (I - vv^T), \nabla H \otimes \nabla H - V\rangle^2 \leq 4\mathbb{E}[\|\nabla H\|^4].$$

The quantity $\mathbb{E}[\|\nabla H\|^4]$ is at most N^2 by the above proved second item in the definition of localizability. This is therefore $O_p(\delta_N^{-2}) = o(\delta_N^{-3})$ as claimed. □

Proof of Theorem 3.2. Having checked localizability for $\mathbf{u}_N$, we apply Theorem 2.3 [4]. To compute $\mathcal{F}$, by the above,

$$\begin{cases}\mathcal{F}_m = 2\mathbb{E}_{a_1,a_2}[a_1 f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))], \\ \mathcal{F}_{r_\perp^2} = 4r_\perp \mathbb{E}_{a_1,a_2}[a_2 f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))].\end{cases}$$

We next turn to calculating the corrector. Recall $V = \mathbb{E}[\nabla H \otimes \nabla H]$, we have that

$$V_{ij} = \mathbb{E}[\partial_i H \partial_j H] = \mathbb{E}[\partial_i L_N \partial_j L_N] - \partial_i \Phi \partial_j \Phi,$$

where

$$\begin{aligned}\mathbb{E}[\partial_i L_N \partial_j L_N] &= 4\mathbb{E}[a_i a_j f'^2(\langle a, x\rangle)(f(\langle a, x\rangle) - f(a_1))^2] + 4\mathbb{E}[\epsilon^2 a_i a_j f'^2(\langle a, x\rangle)] \\ &= 4\mathbb{E}[a_i a_j]\mathbb{E}[f'^2(\langle a, x\rangle)(f(\langle a, x\rangle) - f(a_1))^2] + 4C_\epsilon \mathbb{E}[a_i a_j]\mathbb{E}[f'^2(\langle a, x\rangle)] \\ &+ 4\mathrm{Cov}(a_i a_j, f'^2(\langle a, x\rangle)(f(\langle a, x\rangle) - f(a_1))^2) + 4\mathrm{Cov}(a_i a_j, f'^2(\langle a, x\rangle)).\end{aligned}$$

In particular, for $\delta_N = c_\delta / N$, we have $\delta_N \mathcal{L}_N m = 0$ and

$$\delta_N \mathcal{L}_N r_\perp^2 = \frac{c_\delta}{N} \sum_{i=2}^N V_{ii} = \frac{c_\delta}{N} \sum_{i=2}^N \mathbb{E}[(\partial_i H)^2] = \frac{c_\delta}{N} \sum_{i=2}^N \mathbb{E}[(\partial_i L_N)^2] - \frac{c_\delta}{N} \sum_{i=2}^N (\partial_i \Phi)^2. \tag{5.1}$$

For the first term, one may write

$$\begin{aligned}\sum_{i=2}^N \mathbb{E}[(\partial_i L_N)^2] &= 4(N-1)\left(\mathbb{E}\left[f'^2(\langle a, x\rangle)(f(\langle a, x\rangle) - f(a_1))^2\right] + C_\epsilon \mathbb{E}\left[f'^2(\langle a, x\rangle)\right]\right) \\ &+ 4\sum_{i=2}^N \mathrm{Cov}(a_i^2, f'^2(\langle a, x\rangle)(f(\langle a, x\rangle) - f(a_1))^2) + 4C_\epsilon \sum_{i=2}^N \mathrm{Cov}(a_i^2, f'^2(\langle a, x\rangle)).\end{aligned} \tag{5.2}$$

By the Cauchy-Schwarz inequality,

$$\sum_{i=2}^N \mathrm{Cov}(a_i^2, f'^2(\langle a, x\rangle)(f(\langle a, x\rangle) - f(a_1))^2) \le \sqrt{2(N-1)\mathrm{Var}(f'^2(\langle a, x\rangle)(f(\langle a, x\rangle) - f(a_1))^2)},$$

$$\sum_{i=1}^N \mathrm{Cov}(a_i^2, f'(\langle a, x\rangle)) \le \sqrt{2(N-1)\mathrm{Var}(f^2(\langle a, x\rangle))}.$$

This results in the covariance terms being $O_p(\sqrt{N})$, which vanishes in terms of the correction when it is multiplied by the step-size $\delta_N = c_\delta / N$. Similarly, the population term, $\sum_{i=2}^N (\partial_i \Phi)^2 = 4r_\perp^2 (\partial_{r_\perp^2} \Phi)^2 = O_p(1)$, can be seen to vanish in the corrector as $N \to \infty$.

Finally, the corrector is given by

$$\mathcal{G}_{r_\perp^2} = 4c_\delta \mathbb{E}_{a_1, a_2}\left[f'^2(a_1 m + a_2 r_\perp)\left((f(a_1 m + a_2 r_\perp) - f(a_1))^2 + C_\epsilon\right)\right].$$

Together, these yield the ODE system for $(m, r_\perp^2)$,

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -2\mathbb{E}_{a_1, a_2}[a_1 f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))], \tag{5.3}$$

$$\begin{aligned}\frac{\mathrm{d}r_\perp^2}{\mathrm{d}t} &= -4\mathbb{E}_{a_1, a_2}[a_2 r_\perp f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))] \\ &+ 4c_\delta \mathbb{E}_{a_1, a_2}[f'^2(a_1 m + a_2 r_\perp)((f(a_1 m + a_2 r_\perp) - f(a_1))^2 + C_\epsilon)].\end{aligned} \tag{5.4}$$

Now, let us find the fixed point of m in the above system. Recalling the Hermite polynomials, $(h_k(x))_{k=0}^\infty$, the activation function can be expressed in terms of Hermite polynomials as follows

$$f(x) = \sum_k \alpha_k h_k(x) \quad \text{where,} \quad \alpha_k = \langle f, h_k\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty f(z) h_k(z) e^{-z^2/2} dz.$$

Since $\mathbb{E}[h_k] = 0$ for all $k \geq 1$, we get $\mathbb{E}[f(a_1)] = \alpha_0$. Now, evaluating the ODE at $m = 0$, we obtain the ODE for m given by

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -2\mathbb{E}_{a_1,a_2}[a_1 f'(a_2 r_\perp) f(a_2 r_\perp)] + 2\mathbb{E}_{a_1,a_2}[a_1 f(a_1) f'(a_2 r_\perp)].$$

Then, using Stein's lemma, one may write

$$\frac{\mathrm{d}m}{\mathrm{d}t} = 2\mathbb{E}_{a_1}[f'(a_1)]\mathbb{E}_{a_2}[f'(a_2 r_\perp)]. \tag{5.5}$$

Moreover, $f'(x) = \sum_k \beta_k h_k$ where $\beta_k = (k+1)\alpha_{k+1}$. Hence, $\mathbb{E}[f'(a_1)] = \alpha_1$ and

$$\frac{\mathrm{d}m}{\mathrm{d}t} = 2\alpha_1 \mathbb{E}_{a_2}[f'(a_2 r_\perp)] = 0.$$

□

Proof of Corollary 3.3. Since $m = 0$ is the initial point of the dynamic, one may write

$$\begin{aligned}\frac{\mathrm{d}r_\perp^2}{\mathrm{d}t} =& -4\mathbb{E}_{a_1,a_2}[a_2 r_\perp f'(a_2 r_\perp)(f(a_2 r_\perp) - f(a_1))] \\ &+ 4c_\delta \mathbb{E}_{a_1,a_2}[f'^2(a_2 r_\perp)((f(a_2 r_\perp) - f(a_1))^2 + C_\epsilon)].\end{aligned}$$

Using Stein's lemma, one may write

$$\begin{aligned}\frac{\mathrm{d}r_\perp^2}{\mathrm{d}t} =& 4\mathbb{E}_{a_2}[f'^2(a_2 r_\perp)](c_\delta C_\epsilon + c_\delta \|f\|_{L^2}^2 - r_\perp^2) \\ &+ 4c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp) f^2(a_2 r_\perp)] - 4r_\perp^2 \mathbb{E}_{a_2}[f''(a_2 r_\perp) f(a_2 r_\perp)] \\ &+ 4\mathbb{E}_{a_1}[f(a_1)]\left(r_\perp^2 \mathbb{E}_{a_2}[f''(a_2 r_\perp)] - 2c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp) f(a_2 r_\perp)]\right).\end{aligned} \tag{5.6}$$

□

5.2 Proofs of Theorem 3.5 and Corollary 3.6

Lemma 5.2. *In a single-index model, the distribution of the loss $L_N(x, (a, y))$ depends only on $\tilde{\mathbf{u}}_N = (\sqrt{N}m, r_\perp^2)$. Also, $\tilde{\mathbf{u}}_N$ is δ_N-localizable for E_K being the centered balls of radius K in $\mathbb{R}^2$.*

Proof. By rotational invariance of the Gaussian ensemble, we may take $x^* = v$ where v is the first basis vector of $\mathbb{R}^N$. We have checked localizability in Lemma 5.1, but the change from the original variables is in the J_N matrix, in which now $\nabla\tilde{m} = \sqrt{N}\nabla m = \sqrt{N}v$. This does not affect the first two conditions of localizability; for the third condition, notice that for some $C > 0$

$$\mathbb{E}[\langle \nabla H, \nabla\tilde{m}\rangle^4] = N^2\mathbb{E}[\langle \nabla H, v\rangle^4] = N^2\mathbb{E}[(\partial_1 H)^4] \leq N^2 C,$$

and the second part of the third condition is unchanged since $\nabla^2\tilde{m} = 0$. □

Proof of Theorem 3.5. Most of the bounds assumed in the definition of localizability are used to establish tightness and to ensure that higher-order terms in Taylor expansions vanish in the $N \to \infty$ limit.

Having checked localizability for $\tilde{\mathbf{u}}_N$, in a neighborhood of $m = 0$, by Taylor expansion,

$$f(a_1 m + a_2 r_\perp) = f(a_2 r_\perp) + \frac{\tilde{m}}{\sqrt{N}} a_1 f'(a_2 r_\perp) + \frac{\tilde{m}^2}{2N} a_1^2 f''(a_2 r_\perp) + O_p(N^{-3/2}).$$

Then the population can be expressed as follows

$$\phi(\tilde{m}, r_\perp^2) = \mathbb{E}_{a_1,a_2}[(f(a_2 r_\perp) - f(a_1))^2] + \frac{\tilde{m}^2}{N}\mathbb{E}_{a_2}[f''(a_2 r_\perp) f(a_2 r_\perp)] + \frac{\tilde{m}^2}{N}\mathbb{E}_{a_2}[f'^2(a_2 r_\perp)] - \frac{\tilde{m}^2}{N}\mathbb{E}_{a_1}[a_1^2 f(a_1)]\mathbb{E}_{a_2}[f''(a_2 r_\perp)] + O(N^{-3/2}).$$

Note that, by Stein's lemma $\mathbb{E}_{a_1}[a_1^2 f(a_1)] = \mathbb{E}[f(a_1) + f''(a_1)]$.

One may write the derivatives for $\tilde{\mathbf{u}}_N$ as

$$\nabla\tilde{m} = \sqrt{N}v, \qquad \nabla r_\perp^2 = 2(x - mv).$$

Similar to the standard summary statistics, $\nabla^2\tilde{m} = 0$, while $\nabla^2 r_\perp^2 = 2(I - vv^T)$, and $\nabla^l r_\perp^2 = 0$ for all $l \geq 3$. It yields that

$$\langle\nabla\tilde{m}, \nabla\tilde{m}\rangle = N, \qquad \langle\nabla\tilde{m}, \nabla r_\perp^2\rangle = 0, \qquad \langle\nabla r_\perp^2, \nabla r_\perp^2\rangle = 4r_\perp^2$$

One may apply Theorem 2.3 [4]. To compute $\mathcal{F}$, by the above

$$\begin{cases} \mathcal{F}_{\tilde{m}} = 2\tilde{m}\left(\mathbb{E}_{a_2}[(f'^2(a_2 r_\perp) + f(a_2 r_\perp)f''(a_2 r_\perp)] - E_{a_1}[f(a_1) + f''(a_1)]\mathbb{E}_{a_2}[f''(a_2 r_\perp)]\right), \\ \mathcal{F}_{r_\perp^2} = 4r_\perp \mathbb{E}_{a_2}[a_2 f'(a_2 r_\perp)(f(a_2 r_\perp) - \alpha_0)]. \end{cases}$$

In particular, for $\delta_N = c_\delta/N$, we have $\delta_N \mathcal{L}_N m = 0$. Thus, the corrector $\mathcal{G}_m = 0$. Moreover, in a neighborhood of $m = 0$, the only term that survives in the limit $N \to \infty$ of $\mathcal{G}_{r_\perp^2}$ is the zeroth-order terms of f and f', i.e.,

$$\begin{aligned} \mathcal{G}_{r_\perp^2} &= 4c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp) f^2(a_2 r_\perp)] + 4c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp)]\left(\mathbb{E}_{a_1}[f^2(a_1)] + C_\epsilon\right) \\ &\quad - 8c_\delta \mathbb{E}_{a_1}[f(a_1)]\mathbb{E}_{a_2}[f'^2(a_2 r_\perp) f(a_2 r_\perp)]. \end{aligned}$$

Finally, we consider the volatility of the stochastic process one gets in the limit. Rescaling $J_N V J_N^T$ by noticing that the rescaling $J_N \mapsto \tilde{J}_N$ multiplies the $(1,1)$-entry of $J_N V J_N^T$ by N and its off-diagonal entries by $\sqrt{N}$, one may obtain the entries as follows. Thus,

$$J_N = \begin{pmatrix} \nabla m \\ \nabla r^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 2x_2 & \cdots & 2x_N \end{pmatrix}, \qquad J_N V J_N^T = \begin{pmatrix} V_{11} & 2\sum_{i=2}^N x_i V_{1i} \\ 2\sum_{i=2}^N x_i V_{1i} & 4\sum_{i=2}^N x_i x_j V_{ij} \end{pmatrix}.$$

Again, in a neighborhood of $m = 0$, the only term that survives in the limit $N \to \infty$ of $J_N V J_N^T$ is the zeroth-order terms of f and f'. The (1, 2)-entry of the volatility is given by

$$\begin{aligned} \frac{1}{2}(J_N V J_N^T)_{12} &= \sum_{i=2}^N x_i V_{1i} = \sum_{i=2}^N x_i \mathbb{E}[\partial_1 H \partial_i H] \\ &= 4\sum_{i=2}^N \mathbb{E}_a[a_1 a_i x_i f'^2(\langle a, x\rangle)((f(\langle a, x\rangle) - f(a_1))^2 + C_\epsilon)] \\ &= 4\mathbb{E}_{a_1,a_2}[a_1 a_2 r_\perp f'^2(a_2 r_\perp)((f(a_2 r_\perp) - f(a_1))^2 + C_\epsilon)], \end{aligned}$$

and since the rescaled volatility is $\delta_N(\tilde{J}_N V \tilde{J}_N^T)_{1,2} = \frac{c_\delta}{N}\sqrt{N}(J_N V J_N^T)_{1,2}$, after taking limits, this term will vanish. We could also say the same reason for the (2, 1) entry.

For the (2, 2)-entry, one may write

$$\begin{aligned} \frac{1}{4}(J_N V J_N^T)_{22} &= \sum_{i,j=2}^N x_i x_j V_{ij} = \sum_{i,j=2}^N x_i x_j \mathbb{E}[\partial_i H \partial_j H] \\ &= 4\sum_{i,j=2}^N \mathbb{E}_a[a_i x_i a_j x_j f'^2(\langle a, x\rangle)((f(\langle a, x\rangle) - f(a_1))^2 + C_\epsilon)] \\ &= 4\mathbb{E}_{a_2}[a_2^2 r_\perp^2 f'^2(a_2 r_\perp)((f(a_2 r_\perp) - f(a_1))^2 + C_\epsilon)]. \end{aligned}$$

Again, the rescaled volatility is $\delta_N(\tilde{J}_N V \tilde{J}_N^T)_{2,2} = \frac{c_\delta}{N}(J_N V J_N^T)_{2,2}$ vanishes when $N \to \infty$. It means the only surviving entry of the volatility is the (1, 1)-entry that is given by

$$(J_N V J_N^T)_{11} = V_{11} = \mathbb{E}[(\partial_1 H)^2] = 4\mathbb{E}_{a_1,a_2}[a_1^2 f'^2(a_2 r_\perp)((f(a_2 r_\perp) - f(a_1))^2 + C_\epsilon)].$$

Hence, the volatility is of the form

$$\begin{aligned}\Sigma_{11} = {} & 4c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp)](\mathbb{E}_{a_1}[a_1^2 f^2(a_1)] + C_\epsilon) + 4c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp) f^2(a_2 r_\perp)] \\ & - 8c_\delta \mathbb{E}_{a_1}[a_1^2 f(a_1)]\mathbb{E}[f'^2(a_2 r_\perp) f(a_2 r_\perp)], \qquad \Sigma_{21} = \Sigma_{12} = \Sigma_{22} = 0.\end{aligned}$$

By integration by parts and Stein's lemma, one may write

$$\mathbb{E}_{a_1}[a_1^2 f^2(a_1)] = \mathbb{E}_{a_1}[f^2(a_1) + 2f'^2(a_1) + 2f(a_1) f''(a_1)].$$

Therefore,

$$\begin{aligned}\Sigma_{11} = {} & 4c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp)]\big(\|f\|_{L^2}^2 + 2\|f'\|_{L^2}^2 + 2\langle f, f''\rangle_{L^2} + C_\epsilon\big) \\ & + 4c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp) f^2(a_2 r_\perp)] - 8c_\delta(\mathbb{E}_{a_1}[f(a_1) + f''(a_1)])\mathbb{E}[f'^2(a_2 r_\perp) f(a_2 r_\perp)].\end{aligned}$$

Together, these yield the SDE system for $(\tilde{m}, r_\perp^2)$,

$$\begin{aligned}\mathrm{d}\tilde{m} = {} & -2\tilde{m}\mathbb{E}_{a_2}[f'^2(a_2 r_\perp) + (f(a_2 r_\perp) - (\alpha_0 + \alpha_2)) f''(a_2 r_\perp)]\mathrm{d}t \\ & + 2\sqrt{c_\delta}\big(\mathbb{E}_{a_2}[f'^2(a_2 r_\perp)]\big(\|f\|_{L^2}^2 + 2\|f'\|_{L^2}^2 + 2\langle f, f''\rangle_{L^2} + C_\epsilon\big) \\ & \quad + \mathbb{E}_{a_2}[f'^2(a_2 r_\perp) f^2(a_2 r_\perp)] - 2(\alpha_0 + \alpha_2)\mathbb{E}[f'^2(a_2 r_\perp) f(a_2 r_\perp)]\big)^{1/2}\mathrm{d}B_t, \\ \frac{\mathrm{d}r_\perp^2}{\mathrm{d}t} = {} & 4\mathbb{E}_{a_2}[f'^2(a_2 r_\perp)](c_\delta C_\epsilon + c_\delta\|f\|_{L^2}^2 - r_\perp^2) \\ & + 4c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp) f^2(a_2 r_\perp)] - 4r_\perp^2 \mathbb{E}_{a_2}[f''(a_2 r_\perp) f(a_2 r_\perp)] \\ & + 4\alpha_0\left(r_\perp^2 \mathbb{E}_{a_2}[f''(a_2 r_\perp)] - 2c_\delta \mathbb{E}_{a_2}[f'^2(a_2 r_\perp) f(a_2 r_\perp)]\right).\end{aligned}$$

By (3.1), one may rewrite the dynamics as follows,

$$\begin{aligned}\mathrm{d}\tilde{m} = {} & -2\tilde{m}\left[\mathsf{a}(r_\perp) + \mathsf{b}(r_\perp) - (\alpha_0 + \alpha_2)\mathsf{s}(r_\perp)\right]\mathrm{d}t \\ & + 2\sqrt{c_\delta}\sqrt{\mathsf{q}(r_\perp) - 2(\alpha_0 + \alpha_2)\mathsf{t}(r_\perp) + \mathsf{a}(r_\perp)\big(\|f\|_{L^2}^2 + 2\|f'\|_{L^2}^2 + 2\langle f, f''\rangle_{L^2} + C_\epsilon\big)}\mathrm{d}B_t, \\ \frac{\mathrm{d}r_\perp^2}{\mathrm{d}t} = {} & 4\mathsf{a}(r_\perp)\big(c_\delta C_\epsilon + c_\delta\|f\|_{L^2}^2 - r_\perp^2\big) + 4c_\delta \mathsf{q}(r_\perp) - 4r_\perp^2 \mathsf{b}(r_\perp) + 4\alpha_0\big(r_\perp^2 \mathsf{s}(r_\perp) - 2c_\delta \mathsf{t}(r_\perp)\big).\end{aligned}$$

□

Proof of Corollary 3.6. Since the information exponent of f is at least three, its Hermite coefficients satisfy $\alpha_1 = \alpha_2 = 0$. Substituting $\alpha_2 = 0$, while keeping $\alpha_0 \neq 0$, and evaluating Theorem 3.5 at the fixed point $r_\perp^*$ gives the stated drift $\gamma(r_\perp^*)$. It remains to show $\gamma(r_\perp^*) > 0$.

Recall (3.4), $\gamma(r_\perp^*) = \mathsf{a}(r_\perp^*) + \mathsf{b}(r_\perp^*) - \alpha_0 \mathsf{s}(r_\perp^*)$. If $\alpha_2 = 0$, one may write,

$$\mathsf{a}(r_\perp^*)(c_\delta C_\epsilon + c_\delta\|f\|_{L^2}^2 - r_\perp^{*2}) + \mathsf{q}(r_\perp^*) - r_\perp^{*2}\mathsf{b}(r_\perp^*) + \alpha_0(r_\perp^{*2}\mathsf{s}(r_\perp^*) - 2c_\delta \mathsf{t}(r_\perp^*)) = 0.$$

Solving for $r_\perp^{*2}\mathsf{b}(r_\perp^*)$ and substituting into $\gamma(r_\perp^*)$, we have

$$\gamma(r_\perp^*) = \frac{c_\delta}{r_\perp^{*2}}(\mathsf{a}(r_\perp^*)(C_\epsilon + \|f\|_{L^2}^2) + \mathsf{q}(r_\perp^*) - 2\alpha_0 \mathsf{t}(r_\perp^*)).$$

Moreover,

$$\mathsf{q}(r_\perp^*) - 2\alpha_0 \mathsf{t}(r_\perp^*) = \mathbb{E}_{a_2}[f'^2(a_2 r_\perp^*)(f(a_2 r_\perp^*) - \alpha_0)^2] - \alpha_0^2 \mathsf{a}(r_\perp^*),$$

and using $\|f\|_{L^2}^2 - \alpha_0^2 = \mathbb{E}_{a_1}[(f(a_1) - \alpha_0)^2] \geq 0$, we have

$$\gamma(r_\perp^*) = \frac{1}{r_\perp^{*2}}\big(\mathbb{E}_{a_2}[f'^2(a_2 r_\perp^*)](C_\epsilon + \mathbb{E}_{a_1}[(f(a_1) - \alpha_0)^2]) + \mathbb{E}_{a_2}[f'^2(a_2 r_\perp^*)(f(a_2 r_\perp^*) - \alpha_0)^2]\big). \tag{5.7}$$

On the RHS, the first term is a sum of the noise variance C_ϵ and $\mathrm{Var}(f(a_1))$, multiplied by $\mathbb{E}_{a_2}[f'^2(a_2 r_\perp^*)] \geq 0$, and the second term is positive, because it is an expectation of a product of squares, meaning that $\gamma(r_\perp^*) > 0$. Hence, the rescaled correlation $\tilde{m}$ is mean-reverting to zero near the fixed point $r_\perp^*$. □

References

[1] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A. Erdogdu, *Normal Approximation for Stochastic Gradient Descent via Non-Asymptotic Rates of Martingale CLT*, Proceedings of the Thirty-Second Conference on Learning Theory, PMLR, June 2019, ISSN: 2640-3498, pp. 115–137 (en).

[2] Dyego Araújo, Roberto I. Oliveira, and Daniel Yukimura, *A mean-field limit for certain deep neural networks*, June 2019, arXiv:1906.00193 [math].

[3] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath, *Online stochastic gradient descent on non-convex losses from high-dimensional inference*, Journal of Machine Learning Research **22** (2021), no. 106, 1–51. MR4279757

[4] Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath, *High-dimensional limit theorems for SGD: Effective dynamics and critical scaling*, Communications on Pure and Applied Mathematics **77** (2024), no. 3, 2030–2080 (en), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.22169. MR4692886

[5] Michel Benaïm, *Dynamics of stochastic approximation algorithms*, Séminaire de Probabilités XXXIII (Berlin, Heidelberg) (Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor, eds.), Springer, 1999, pp. 1–68 (en). MR1767993

[6] Albert Benveniste, Michel Métivier, and Pierre Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer, Berlin, Heidelberg, 1990. MR1082341

[7] M. Biehl and H. Schwarze, *Learning by on-line gradient descent*, Journal of Physics A: Mathematical and General **28** (1995), no. 3, 643 (en). MR1326097

[8] Léon Bottou, *On-line Learning and Stochastic Approximations*, On-Line Learning in Neural Networks (David Saad, ed.), Publications of the Newton Institute, Cambridge University Press, Cambridge, 1999, pp. 9–42. MR1763695

[9] Michael Celentano, Chen Cheng, and Andrea Montanari, *The high-dimensional asymptotics of first order methods with random data*, April 2026, arXiv:2112.07572 [math].

[10] Lénaïc Chizat and Francis Bach, *On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport*, Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018.

[11] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi, *Hitting the High-dimensional notes: an ODE for SGD learning dynamics on GLMs and multi-index models*, Information and Inference: A Journal of the IMA **13** (2024), no. 4, iaae028 (en). MR4811680

[12] Paul Dupuis and Harold J. Kushner, *Stochastic Approximation and Large Deviations: Upper Bounds and w.p.1 Convergence*, SIAM Journal on Control and Optimization **27** (1989), no. 5, 1108–1135, Publisher: Society for Industrial and Applied Mathematics. MR1009340

[13] Sebastian Goldt, Madhu Advani, Andrew Saxe, Florent Krzakala, and Lenka Zdeborová, *Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup*, Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019. MR4241363

[14] Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa, *Tight analyses for non-smooth stochastic gradient descent*, Proceedings of the Thirty-Second Conference on Learning Theory, PMLR, June 2019, ISSN: 2640-3498, pp. 1579–1613 (en).

[15] Harold J. Kushner, *Asymptotic behavior of stochastic approximation and large deviations*, The 22nd IEEE Conference on Decision and Control, December 1983, pp. 75–81. MR0764693

[16] Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang, *Diffusion Approximations for Online Principal Component Estimation and Global Convergence*, Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.

[17] Qianxiao Li, Cheng Tai, and Weinan E, *Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations*, Journal of Machine Learning Research **20** (2019), no. 40, 1–47. MR3948080

[18] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora, *On the Validity of Modeling SGD with Stochastic Differential Equations (SDEs)*, Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 12712–12725.

[19] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Transactions on Automatic Control **22** (1977), no. 4, 551–575. MR0465458

[20] Stephan Mandt, Matthew D. Hoffman, and David M. Blei, *Stochastic Gradient Descent as Approximate Bayesian Inference*, Journal of Machine Learning Research **18** (2017), no. 134, 1–35. MR3763768

[21] D. L. McLeish, *Functional and random central limit theorems for the Robbins-Munro process*, Journal of Applied Probability **13** (1976), no. 1, 148–154 (en). MR0403116

[22] Song Mei, Andrea Montanari, and Phan-Minh Nguyen, *A mean field view of the landscape of two-layer neural networks*, Proceedings of the National Academy of Sciences **115** (2018), no. 33, E7665–E7671, Publisher: Proceedings of the National Academy of Sciences. MR3845070

[23] Deanna Needell, Nathan Srebro, and Rachel Ward, *Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm*, Advances in Neural Information Processing Systems, vol. 27, Curran Associates, Inc., 2014. MR3439812

[24] Herbert Robbins and Sutton Monro, *A Stochastic Approximation Method*, The Annals of Mathematical Statistics **22** (1951), no. 3, 400–407 (en), Publisher: Institute of Mathematical Statistics. MR0042668

[25] Grant Rotskoff and Eric Vanden-Eijnden, *Trainability and Accuracy of Artificial Neural Networks: An Interacting Particle System Approach*, Communications on Pure and Applied Mathematics **75** (2022), no. 9, 1889–1935 (en), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.22074. MR4465905

[26] David Saad and Sara A. Solla, *Exact Solution for On-Line Learning in Multilayer Neural Networks*, Physical Review Letters **74** (1995), no. 21, 4337–4340, Publisher: American Physical Society.

[27] Justin Sirignano and Konstantinos Spiliopoulos, *Mean field analysis of neural networks: A central limit theorem*, Stochastic Processes and their Applications **130** (2020), no. 3, 1820–1852. MR4058290

[28] Yan Shuo Tan and Roman Vershynin, *Online Stochastic Gradient Descent with Arbitrary Initialization Solves Non-smooth, Non-convex Phase Retrieval*, October 2019, arXiv:1910.12837 [stat]. MR4582480

[29] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová, *Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks*, Advances in Neural Information Processing Systems **35** (2022), 23244–23255 (en). MR4725421

Acknowledgments. I would like to thank Aukosh Jagannath, who supervised my master's thesis at the University of Waterloo, for his discussions and resources; his guidance influenced both that thesis and this paper.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS[1])
- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4], ProjectEuclid[5]
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[2]EJMS: Electronic Journal Management System: `https://vtex.lt/services/ejms-peer-review/`
[3]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[4]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[5]Project Euclid: `https://projecteuclid.org/`
[6]IMS Open Access Fund: `https://imstat.org/shop/donation/`